Scholarly Citation Prediction Using Altmetrics

Pavan Ravikanth Kondamudi
Northern Illinois University
DeKalb, Illinois
pkondamudi@niu.edu

ABSTRACT

Given the extraordinary growth in the quantity of scholarly literature published annually and the potential of some to have a significant effect on many aspects of our lives, identifying at an early stage meaningful scholarly work with the potential to have the greatest impact on society is a vital endeavor. Identifying such work is of great importance to the academic research community and to other stakeholders, such as technology companies and government bodies. Given the sheer amount of research published and the growth of ever-changing interdisciplinary areas, researchers need an effective approach to identifying important scholarly studies if they are to read or even skim all the new studies published in their respective fields. The number of citations in scholarly work that a given research article has accrued and the rank of the scholarly venues in which the work has been cited have been used to help researchers in this regard. However, citations take time to occur and longer to accumulate. In the present paper, we build classification models to predict whether or not a research article will be cited in a scholarly work using non-traditional features (altmetrics) and context features of social media posts that are related to altmetrics. We found that the tree-based models, namely, Random Forest and Decision Tree, outperformed the other models we tested.

KEYWORDS

Altmetrics, Scientometrics, Scholarly Communication, Social Media

1 INTRODUCTION

Every year, a massive amount of research in all fields of study is conducted worldwide. To filter out articles and determine their scholarly impact, researchers have used a number of approaches, among which are consulting peer reviews, conducting a citation analysis, and assessing the text of an article [10]. However, most of these approaches are time-consuming. And, because they ignore many other avenues through which research receives attention, they are also self-limiting. An early and accurate identification of articles with the potential to make a significant impact is, therefore, necessary. An effective approach of this nature would provide a basis for stakeholders to explore research in a timely way such that it would be possible to identify important results reported in sound research articles and act on them more quickly than is presently the case. Thus, research would have an impact as quickly as possible.

CHIIR 18, March 2018, New Brunswick, NJ, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Citations are given great importance in efforts to measure the impact of scholarly works. Yet, in general, following a traditional model of publication, it takes months or even years for a paper to be cited, as research takes a long journey both to publication and then subsequently to citation. This journey involves a number of steps: other researchers must discover the article, then peer review it, then it must be published, cited in another manuscript, which must then go through the same process before it is available to the scholarly community. Researchers often go through high volumes of articles to determine best relevancy, taking time from the more creative aspects of research. Filtering out low-quality work from the research process is critical, therefore, in order to save time on an individual basis and to accelerate research and its impact on a larger scale. Given the length of time it takes for an article to be published and for citations to accrue and given the proliferation of research and of online platforms where it is made available, there is a clear need across multiple fields for a more efficient way to evaluate the quality of research articles. Thus, alternative approaches are needed.

The advent of online platforms such as social media sites and advances in data modeling capabilities make it possible to capture complex trends in data pertaining to multiple areas, political, civic, scientific, cultural, educational, and economic, etc., and to predict trends and events. As online platforms serve as a means of informal discussion, researchers are increasingly using them for scientific and professional communications. This trend in the online context presents opportunities to identify quality research, to find experts, and to identify potential societal impact in faster, more comprehensive, and more diverse ways.

Article-level metrics (altmetrics) have been proposed as both an alternative and a complement to established scholarly metrics such as citations, the Impact Factor (IF), the immediacy index, and the h-index. Altmetrics is a growing area of interest that intends to measure the societal impact of research based on the dissemination of a research outcome on multiple social media platforms such as Facebook and Twitter, reference managers such as Mendeley and CiteULike, and information websites such as Wikipedia, online news outlets, blogs, and other peer review websites. Further, some websites such as Altmetric.com, ImpactStory.org, and Paperbuzz.org collect altmetrics data for various research articles.

Several research communities pursue citation count prediction in order to locate important research in its early stages. In many studies, researchers have considered numerous data relating to measuring the quality and impact of research papers, such as textual features, the number of times it has been downloaded, and the author's h-index. In the present paper, we investigate altmetrics and the content features of social media posts related to altmetrics as indicators of whether or not an article will receive at least one scholarly citation.

2 RELATED WORK

Several studies have been conducted to measure and predict the scholarly impact of research articles. Manjunatha et al. [8] and Davletov et al. [2] used various machine learning techniques to consider temporal features in predicting the citation performance of papers. In addition to machine learning models, statistical techniques have also been applied. For instance, the Ordered Probit model [9], a statistical model used by Perlich et al. [13], and a spatio-temporal function [15] have each been used to predict a paper's citation counts. Ismail and Mohadn used social signals [1] to rank the importance of a document based on a consideration of temporal aspects.

New metrics for citation prediction, namely future influence prediction and citation count prediction, have been proposed by Yan et al. [19][20], who considered a range of features including authors, content, venues, topic rank, diversity, h-index, author rank, productivity, sociality, authority, venue rank, and venue centrality. Based on their evaluation of several prediction models, i.e., the Linear Regression, k-Nearest Neighbor (kNN), Support Vector Regression (SVR), and CART models, they found that the authors of scholarly articles evince prejudice in their citation practices. They propose that the author's expertise and the impact of the venue are the crucial features in determining a paper's citation count. Similarly, Hirsch [5] found that the h-index is an important feature in predicting the future achievement of a scholarly work. He made a comparison with other features such as the number of citations, the number of papers, and the mean number of citations per paper. Nezhadbiglari et al. [12] computed the popularity trend using the spectral clustering algorithm based on the total citation count as a measure of scientific popularity. By extracting a set of academic features such as the number of publications a scholar has and the number of venues in which he/she has published, the computed popularity trend is effective in predicting the early fame of a scholar. Harnad and Brody [4] studied the downloads feature as a metric to predict future citations and found a significant correlation between the number of downloads and the number of citations.

In studies on joint modeling of texts, Nallapati et al. [11] analyzed the Pairwise-Link-LDA and the Link-PLSA-LDA models for citation prediction. Of the two models, the researchers observed that the Link-PLSA-LDA model performs better on the citation prediction task. A new predictive model called "term bucketing" in heterogeneous networks, proposed by Yu et al. [21], can capture document and topic similarities without breaking possible citation relations, and put papers into different buckets, thereby helping to reduce the search space for both model learning and for answering citation queries. To interpret hidden network information in a bibliographic dataset, a meta path based feature setup space has been used, which may even be capable of defining citation probability. In related work, Sun et al. [18] used a similar meta-path model along with a time-prediction model to deliver not only citation predictions but also the timeframe within which they would accrue. The countX feature (the number of times a paper is cited in the same article) and citeWords (the number of words within the citation context) are extracted from the citation context and evaluated by Singh et al. [17]. The authors showed that this approach is around 10% more accurate in terms of prediction on average than the nearest baseline.

Kwak and Lee [7] found that articles shared extensively on Twitter are not only from highly ranked journals but also from lower ranked journals. Based on their results, the researchers inferred that publication in a higher ranking journal does not lead to a study having more social impact than one published in a lower ranking journal. A new metric referred to as "twimpact" was proposed by Eysenbach [3], who showed that a tweet can predict whether an article will be highly influential in the three days immediately following its publication. The study results suggest that the twimpact factor can be used to measure and filter research findings.

Sebastian [16] introduced the Literature Based Discovery (LBD) algorithm, which uses computational algorithms to discover potential hidden connections between previously disconnected sets of literature, and Latent Domain Similarity (LDS), which uses combinations of semantic features (e.g., the distribution of technical terms in titles and abstracts) and structural features (e.g., cited references, citing articles) of two or more articles in order to infer shared latent domains between them. The goal of this algorithm is to explore whether these shared latent domains correlate with the probability of previously disconnected clusters to form future citation links with each other. As two sets of literatures could have been published separately in two seemingly unrelated fields, it is possible that they share some or even many domains previously unknown to researchers in each field. Kunegis et al. [6] proposed a link prediction algorithm based on a spectral evolution model, which states that the growth can be described by a change in the spectrum. They studied the comparison of the graph kernel function (a variety of link prediction algorithms) and found that the spectral evolution model provides a justification for more complex link prediction methods. Based on a classification of data from the Brainly educational community question answering site, Rath et al. [14] effectively evaluated a question's quality. Using binary classification, Zoller et al. [22] drew on logs from the BibSonomy social tagging system to predict whether a paper would receive more than a median number of citations in the year following publication. They also found that posts, exports, and visits of publications are correlated with citations.

Overall, the literature shows that in endeavors to predict the scholarly impact of research papers, researchers have considered several approaches and features such as metadata features (e.g., author, affiliation, venue, and publication year), textual features (e.g., abstract and scholarly text), and web-based features (e.g., downloads and views). In the present research study, we address this problem using a variety of models and a combination of altmetrics data and content features.

3 DATASET AND METHODS

We used data from Altmetric.com. The initial dataset consisted of altmetrics for 400,000 research articles. We randomly selected a sample set of 118,042 articles from the original dataset. We used the data to collect updated altmetrics data using Altmetric API (https://api.altmetric.com/) for the selected sample articles using the Digital Document Identifier (DOI) of each. We then mapped the articles in our dataset to articles on Google Scholar by conducting an article title search, and we collected citation counts using a custom written scraping tool (Figure 1).

Figure 1: Citation counts scraped from Google Scholar.

For each research article, we considered two sets of features. First, we considered altmetrics features such as mentions and reads on various online platforms and reference manager websites (Table 1). Then, we considered the content related features of posts on social media platforms that generated altmetrics (Table 2). The content-related features include (1) the post length, which measures the length of a post excluding the paper title and hyperlinks in the post, (2) the number of unique hashtags used in the posts, which indicates the reachability of an article in general terms, and (3) the presence of a link to the article in the post, which renders an article immediately accessible. In building the classification models, we considered eight altmetrics features (Table 1) and three content features (Table 2) for a total of eleven features.

Table 1: Altmetrics Features

Feature	Description
Blogs	Number of mentions in blogs
CiteULike	Number of readers on CiteULike
Google Plus	Number of mentions on Google Plus
Mendeley	Number of readers on Mendeley
Peer reviews	Number of mentions on peer review websites
Policy	Number of citations in policy documents
Total accounts	Number of accounts that posted an article
Twitter	Number of mentions on Twitter

Table 2: Content Features

Feature	Description
Hashtag	Number of unique hashtags used in a post
Link	Presence of a link to the article
Post length	Post length excluding article title and links

We labeled our data using binary indicators. Articles that received one or more citations were assigned a class label "YES" and articles that had not received any citations were assigned a "NO" label. In our dataset, we also had a class imbalance issue with 102,862 observations belonging to the "YES" class and 15,180 observations belonging to "NO" class. To address with this issue, we trained and tested our models using down-sampling and up-sampling strategies.

We split the data according to a 70:30 ratio for training and testing, respectively. Our experiments included a 10-fold cross-validation from which we took the average accuracy score in order to obtain the accuracy scores for the final model. For classification, we trained and tested nine algorithms on our data: the K-Nearest Neighbors, Decision Tree, Random Forest, Neural Net, AdaBoost,

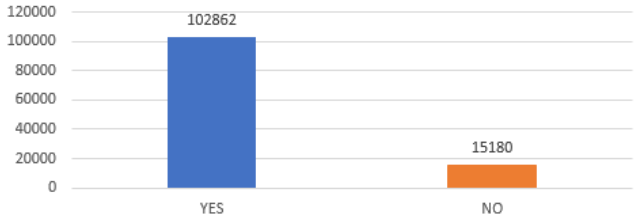

Figure 2: Class Imbalance.

Gaussian, Multinomial, Bernoulli Naive Bayes, and Quadratic Discriminant Analysis (QDA) classifiers. In order to compare the performance of the classifiers, we obtained the accuracy, precision, recall, and F-1 scores. In our experiment, Multilayer Perceptron (MLP) was equipped with 100 units in each of 10 hidden layers, rectified linear unit "ReLU", "L-BFGS" optimizer that belongs to quasi-Newton methods family.

Figure 3 shows a plot between the neighborhood size (x-axis) ranging from 3 to 75 and an average accuracy score over 10-fold cross-validation, precision, recall, and F-1 scores (y-axis) in order to choose the best K value for the KNN classifier. From the plot, it can be observed that the scores were too high where K < 10 but more stable where K >= 10.

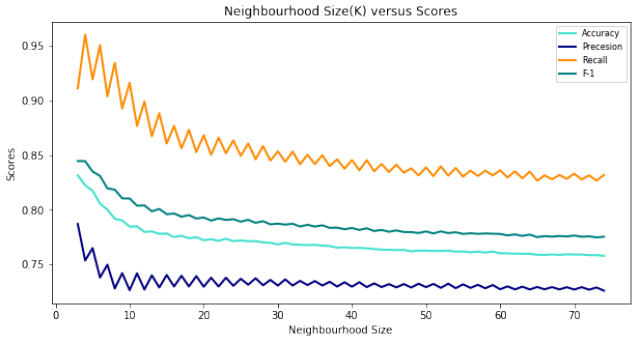

Figure 3: KNN Classifier Neighborhood Convergence.

4 RESULTS

The accuracy, precision, recall, F-1, and area under the ROC curves for up-sampling are shown in Table 3 . Of the models we tried, the tree-based machine-learning algorithms performed best. Further, the up-sampling strategy worked better than did the down-sampling strategy. The presence of more observations helped the models to scale up in terms of accuracy, precision, and recall. None of the models yielded good results with the down-sampling strategy.

To determine which features are most important to yielding accurate predictions, we calculated weights to represent the contribution of each feature. The higher the weight of a feature, the more important it is for classification. The weight of each feature varied across the models. Table 4 shows the importance of each feature from most to least important for the top-performing algorithms, i.e., the Random Forest and Decision Tree models. It can be observed that Mendeley reads and post length were the first and

Table 3: Classifier, Accuracy (A), Precision (P), Recall(R), F-1 score (F1), AUC - ROC (AR)

Classifier	*A*	*P*	*R*	*F1*	*AR*
AdaBoost	.76	.73	.81	.77	.76
Bernoulli Naive Bayes	.62	.58	.87	.63	.62
Decision Tree	.87	.82	.96	.88	.88
Gaussian Naive Bayes	.57	.54	.96	.58	.58
MLP	.71	.69	.87	.77	.75
Multinomial Naive Bayes	.70	.67	.78	.70	.70
KNN (K=10)	.77	.72	.91	.81	.78
QDA	.58	.54	.96	.69	.58
Random Forest	.87	.82	.96	.88	.88

second most important features, respectively, whereas mentions on Google Plus and policy citations were least and second to least important, respectively.

Table 4: Feature Importance Table for the Random Forest and Decision Tree Models in Up-Sampling

Feature	Random Forest	Decision Tree
Mendeley	.42688	.45264
Post length	.37372	.32603
Twitter	.04726	.05568
Hashtags	.04506	.05516
Total accounts	.04063	.05357
Blogs	.00897	.01097
CiteULike	.02341	.00957
Link	.01038	.01286
Peer review	.01158	.01527
Policy	.00653	.00323
Google Plus	.00553	.00499

5 CONCLUSION AND FUTURE WORK

Given the proliferation of research publications, predicting the scholarly impact of research at an early stage would save the scholarly community, research agencies, and policy makers crucial time and thereby accelerate overall research progress. In the present study, we investigated the possibility of using social media features to predict whether a research paper will receive at least one citation. We built and tested several classifiers using altmetrics data and found that tree-based machine learning models performed better than the other models in terms of accuracy, precision, recall, and F-1 scores. In our future research, we intend to work on citation prediction based on deep learning algorithms with more altmetrics features and textual features extracted from research articles.

REFERENCES

[1] Ismail Badache and Mohand Boughanem. 2017. Fresh and Diverse Social Signals: Any Impacts on Search?. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, New York, NY, USA, 155–164. https://doi.org/10.1145/3020165.3020177

[2] Feruz Davletov, Ali Selman Aydin, and Ali Cakmak. 2014. High Impact Academic Paper Prediction Using Temporal and Topological Features. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14* (2014). https://doi.org/10.1145/2661829.2662066

[3] Gunther Eysenbach. 2011. Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of Medical Internet Research* 13, 4 (2011). https://doi.org/10.2196/jmir.2012

[4] Stevan Harnad and Tim Brody. 2004. Prior evidence that downloads predict citations. (2004).

[5] J. E. Hirsch. 2007. Does the h index have predictive power? *Proceedings of the National Academy of Sciences* 104, 49 (2007), 19193–19198. https://doi.org/10.1073/pnas.0707962104

[6] JÃlrÃťme Kunegis, Damien Fay, and Christian Bauckhage. 2010. Network growth and the spectral evolution model. *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10* (2010). https://doi.org/10.1145/1871437.1871533

[7] Haewoon Kwak and Jong Gun Lee. 2014. Has much potential but biased. *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion* (2014). https://doi.org/10.1145/2567948.2576956

[8] J. N. Manjunatha, K. R. Sivaramakrishnan, Raghavendra Kumar Pandey, and M Narasimha Murthy. 2003. Citation prediction using time series approach KDD Cup 2003 (task 1). *ACM SIGKDD Explorations Newsletter* 5, 2 (Jan 2003), 152. https://doi.org/10.1145/980972.980993

[9] Richard D. Mckelvey and William Zavoina. 1975. A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology* 4, 1 (1975), 103–120. https://doi.org/10.1080/0022250x.1975.9989847

[10] Kathy McKeown, Hal Daume, Snigdha Chaturvedi, John Paparrizos, Kapil Thadani, Pablo Barrio, Or Biran, Suvarna Bothe, Michael Collins, Kenneth R. Fleischmann, Luis Gravano, Rahul Jha, Ben King, Kevin McInerney, Taesun Moon, Arvind Neelakantan, Diarmuid O'Seaghdha, Dragomir Radev, Clay Templeton, and Simone Teufel. 2016. Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology* 67, 11 (2016), 2684–2696. https://doi.org/10.1002/asi.23612

[11] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* (2008). https://doi.org/10.1145/1401890.1401957

[12] Masoumeh Nezhadbiglari, Marcos AndrÃĬ GonÃğalves, and Jussara M. Almeida. 2016. Early Prediction of Scholar Popularity. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16* (2016). https://doi.org/10.1145/2910896.2910905

[13] Claudia Perlich, Foster Provost, and Sofus Macskassy. 2003. Predicting citation rates for physics papers. *ACM SIGKDD Explorations Newsletter* 5, 2 (Jan 2003), 154. https://doi.org/10.1145/980972.980994

[14] Manasa Rath, Long T. Le, and Chirag Shah. 2017. Discerning the Quality of Questions in Educational Q&Ausing Textual Features. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval (CHIIR '17)*. ACM, New York, NY, USA, 329–332. https://doi.org/10.1145/3020165.3022145

[15] Peter Z. Revesz. 2014. A method for predicting citations to the scientific publications of individual researchers. *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS '14* (2014). https://doi.org/10.1145/2628194.2628210

[16] Yakub Sebastian. 2014. Cluster links prediction for literature based discovery using latent structure and semantic features. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14* (2014). https://doi.org/10.1145/2600428.2610376

[17] Mayank Singh, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. 2015. The Role Of Citation Context In Predicting Long-Term Citation Profiles. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15* (2015). https://doi.org/10.1145/2806416.2806566

[18] Yizhou Sun, Jiawei Han, Charu C. Aggarwal, and Nitesh V. Chawla. 2012. When will it happen? *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12* (2012). https://doi.org/10.1145/2124295.2124373

[19] Rui Yan, Congrui Huang, Jie Tang, Yan Zhang, and Xiaoming Li. 2012. To better stand on the shoulder of giants. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries - JCDL '12* (2012). https://doi.org/10.1145/2232817.2232831

[20] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction. *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11* (2011). https://doi.org/10.1145/2063576.2063757

[21] Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. 2012. Citation Prediction in Heterogeneous Bibliographic Networks. *Proceedings of the 2012 SIAM International Conference on Data Mining* (2012), 1119–1130. https://doi.org/10.1137/1.9781611972825.96

[22] Daniel Zoller, Stephan Doerfel, Robert Jäschke, Gerd Stumme, and Andreas Hotho. 2016. Posted, visited, exported: Altmetrics in the social tagging system BibSonomy. *Journal of Informetrics* 10, 3 (2016), 732–749.